Stochastic Programming with Probability Constraints

Laetitia Andrieu, Guy Cohen, Felisa Vázquez-Abad

Stochastic Programming with Probability Constraints

Laetitia Andrieu*, Guy Cohen†, Felisa J. Vázquez-Abad‡

July 31, 2007

Abstract

In this work we study optimization problems subject to a failure constraint. This constraint is expressed in terms of a condition that causes failure, representing a physical or technical breakdown. We formulate the problem in terms of a probability constraint, where the level of "confidence" is a modelling parameter and has the interpretation that the probability of failure should not exceed that level. Application of the stochastic Arrow-Hurwicz algorithm poses two difficulties: one is structural and arises from the lack of convexity of the probability constraint, and the other is the estimation of the gradient of the probability constraint. We develop two gradient estimators with decreasing bias via a convolution method and a finite difference technique, respectively, and we provide a full analysis of convergence of the algorithms. Convergence results are used to tune the parameters of the numerical algorithms in order to achieve best convergence rates, and numerical results are included via an example of application in finance.

Keywords. Probability constraints, stochastic programming, stochastic gradient algorithm, stochastic approximation

1 Introduction

1.1 Constrained Optimization in a Stochastic Setting

Optimization Theory provides a convenient approach to formulate and solve problems involving conflicting objectives, which is generally the challenge present in decision making situations. The main idea is to aggregate as many objectives as possible into a single objective function, which may be straightforward when those objectives are amenable to an expression into a common unit, say, a currency unit as dollar or euro. In this objective aggregation, weights are allocated to each term in order to reflect preferences or priorities. However, there might be other objectives that can hardly be expressed in a unit commensurable with the previous ones (examples to come hereafter). In such a case, it is better to introduce those other objectives through constraints, that is, each such objective should not exceed a prescribed level. The constraint levels are set a priori, as are the weights for the different terms in the cost function.

Duality Theory provides the tools to evaluate the sensitivity of the optimal solution (cost) to those prescribed constraint levels. In mathematical terms, let u be the decision variable in a Hilbert space $\mathcal{U}$, $J : \mathcal{U} \to \mathbb{R}$ the cost function, and $\Theta : \mathcal{U} \to \mathbb{R}^d$ the constraint function. We consider problems of the type:

$$\min_{u \in U^{\rm ad}} J(u) \quad \text{s.t.} \quad \Theta(u) \leq \alpha \ , \tag{1}$$

*EDF R&D, Dép. OSIRIS,1 avenue du Général de Gaulle, 92141 Clamart Cedex, France

†CERMICS-ENPC, 6–8 avenue Blaise Pascal, 77455 Marne la Valle Cedex 2, France, guy.cohen@mail.enpc.fr

‡Dept. Math. & Stat., University of Melbourne, 3010 Victoria, Australia

where $U^{\rm ad}$ is an "admissible" or "feasible" closed convex subset of $\mathcal{U}$ and inequalities in the constraints involving Θ are understood componentwise. Introduce the multiplier λ (in $\mathbb{R}^d_+$) and the Lagrangian

$$L(u,\lambda) = J(u) + \langle \lambda, \Theta(u) - \alpha \rangle \ , \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product. Kuhn-Tucker optimality conditions characterize an optimal multiplier $\lambda^\sharp$ which can be interpreted as the sensitivity of the optimal cost function $J(u^\sharp)$ (where $u^\sharp$ denotes the solution of problem (1)) with respect to α (up to a change of sign).

When random factors affect the outcomes of a decision, a classical approach is to assume that the probability distribution of those factors is known and to appeal to *stochastic* optimization. Call ξ the corresponding random variable, then the objective function is usually expressed in terms of an expectation of some cost function of the form $J(u) = \mathbb{E}\big(j(u,\xi)\big)$.

In the stochastic situation, modelling choices for aggregation of objectives, weights and constraints are similar to the deterministic case. However a new question also arises regarding constraints, namely, constraints can be formulated in various ways: "almost surely", "in expectation", "in probability", etc.

The first possibility ("almost sure" constraints) means that certain quantities $\theta(u,\xi)$ depending on decision variables and affected by random factors should satisfy equality or inequality for "almost all" values of those random factors (according to their probability distributions). This is in particular the case of constraints which express "laws of nature" which are part of the mathematical model of the system under consideration. However, regarding objectives or "wishes", such strict constraints are generally inappropriate from the economic or simply realistic point of view. Suppose for example that a pressure should not exceed a certain level beyond which death will almost surely happen. First of all, observe that it is hard to aggregate such an objective (actually, that to stay alive) with other more economic objectives which aim at saving money. Second, under the constraint that the pressure "almost never" exceeds the dangerous level, the operation can be extremely costly if not simply impossible. That is, some *risk* must be accepted for the operation to be economically viable.

The second possibility (constraints "in expectation") means that, given a decision, the *expected value* $\Theta(u) = \mathbb{E}\big(\theta(u,\xi)\big)$ of a critical quantity (a pressure in our example) should not exceed a certain level. Such a formulation is generally mathematically attractive, but it is difficult to understand how much risk is involved in choosing such or such prescribed level. Indeed, given a decision u, the pressure (to keep on with our example) becomes a random variable $\theta(u,\xi)$ with a certain distribution (which is affected by the chosen decision), and the only thing one ask is that the first moment (the expectation) of this random variable stay below a prescribed level, but with no direct control on how much of the probability mass will lie beyond that prescribed level.

The third possibility advocated to (constraints "in probability" or "probabilistic constraints") means that one accepts that the critical quantity (the pressure, say) remains under the prescribed level not "almost always" as earlier, but with a certain probability whose value must be chosen. In mathematical terms, one now considers the problem

$$\min_{u \in U^{\rm ad}} \mathbb{E}\big(j(u,\xi)\big) \quad \text{s.t.} \quad \mathbb{P}\big(\theta(u,\xi) \le \alpha\big) \ge \pi \ . \tag{3}$$

This chosen probability value π exactly reflects the risk one is ready to assume (in contrast with the previous approach of constraints in expectation). As discussed earlier, duality should then help in evaluating the sensitivity of the optimal cost function with respect to this accepted, but arbitrarily fixed, level of risk.

1.2 Quantitative vs. Qualitative Risk Measures

We now motivate the interest of probability constraints in contrast with other measures of risk. Before choosing a risk measure, it is very important to know which type of failure we are interested in: quantitative failure or qualitative failure. For example, a power supply company would minimize its cost under the constraint of supplying the demand. If that demand cannot be fully supplied, it matters to know which percentage of it will not be covered and during which amount of time. This is what we mean by "quantitative failure": introducing a penalty for the total amount of demand not supplied directly into the cost function, or choosing to constrain a quantity which accounts for the amount of supply failure is appropriate in that situation. On the contrary, when simply going beyond a certain threshold causes death, it not does matter to know by which amount that threshold has been exceeded — this is what we mean by "qualitative failure" — but it does matter to know the likelihood of going beyond that critical threshold. Probability constraints are particularly adapted to this latter situation.

In fact, because of the mathematical difficulties raised by probability constraints, these constraints must be exclusively used in the case of qualitative failure problems. For quantitative failure problems, there are other risk measures with better mathematical properties (e.g. convexity), like Conditional Value-at-Risk (CVaR) for instance. Introduced in [13], CVaR is one of the most popular risk measure in finance. CVaR is the average of a random variable for the worst scenarios. Denote by $\alpha_u(\pi)$ the quantile function of the distribution of $\theta(u,\cdot)$ with confidence level π (also called Value-at-Risk). Then, CVaR, denoted by $\phi_\pi(u)$, is defined by

$$\phi_\pi(u) = \mathbb{E}\big(\theta(u,\xi) \mid \theta(u,\xi) \geq \alpha_u(\pi)\big)\,.$$

The risk constraint will be then $\phi_\pi(u) \leq \overline{\alpha}$, where $\overline{\alpha}$ represents the accepted level of risk, and π is fixed a priori.

Notice that the critical threshold α in the probability constraint is generally provided by technical considerations, whereas π characterizes the level of risk one is ready to accept. That is, the decision maker may bargain about the constraint level π but not on that threshold α which is a technical data. With the CVaR approach, this α disappears from the formulation and we believe that this is a weakness of this approach. Moreover, in the case of "qualitative failure", there is no meaning in averaging values of θ beyond a threshold which is supposed to be fatal.

1.3 About this Paper

Problem (3) is the class of problems considered in this paper. Its advantage is again the fact that the meaning of constraints in terms of risk assumed is of immediate perception. Its drawback is its mathematical difficulty.

In this paper, we discuss an approach relying upon Lagrangian duality and stochastic gradient to solve (3). The use of stochastic gradient is based on the reformulation of constraints in probability as constraints in expectation, using an indicator function. As usual with stochastic gradient, we assume that the functions involved in the problem (here, j and θ) are known explicitly but that the probability law governing the "noise" ξ is not, or that the computation of expectations of the variables involved is out of reach or too costly. It is rather assumed that an external mechanism delivers samples of ξ which are used in the iterative algorithm.

Writing the probability as an expectation opens the possibility of using stochastic gradient algorithms, but it also raises the difficulty of handling a discontinuous function, namely the indicator function. We will discuss various ways of overcoming that difficulty.

The rest of the paper is organized as follows. In §2, we present the analysis of the problem, and our resolution strategy, a stochastic Arrow-Hurwicz algorithm. In §3, we describe two

structural difficulties of stochastic programming under probability constraint. To implement a stochastic Arrow-Hurwicz algorithm, we need to handle the probability function gradient. In §4, the question we are interested in is therefore: how to compute stochastic estimates of the probability function gradient? In order to answer this question, we propose two methods that allow to obtain *biased* stochastic gradient estimates, namely Approximation by Convolution (AC) ad Finite Differences (FD). We defer to a forthcoming paper to propose techniques based on integration by parts ideas and providing *unbiased* (or *consistent*) estimates, and to compare them with the biased estimates studied hereafter.

We consider a very basic portfolio optimization problem under a probability constraint and use this example throughout the rest of the paper to illustrate and compare the AC and FD techniques. Section 5 is devoted to the convergence analysis of the proposed methods. Finally, §6 reports numerical experiments with the Arrow-Hurwicz algorithm.

2 Analysis of the Problem

2.1 Review of Main Difficulties

Probability constraints provide a straightforward risk formulation with an immediate intuitive interpretation. But at the same time, it is well known that such constraints raise important mathematical difficulties, such as the lack of convexity or connectedness of the feasible subset. Indeed, even if θ is a convex function of u for almost all values ξ, the constraint in (3) may not define a convex feasible subset in $\mathcal{U}$ (which can even be not connected, if not empty). Those convexity or connectedness (or emptiness) properties depend of course on the properties of θ as a function of its two arguments u and ξ, on the probability distribution of the random variable ξ, on the level α of constraint required and on the level π of probability required. One may refer to [9] for a discussion on those convexity properties, and to [8] for connectedness properties.

In [9], the authors prove that if $\theta(\cdot,\cdot)$ is jointly convex in (u,ξ) and the probability measure is quasi-concave, then the feasible subset of (3) is convex. But those assumptions seem to us to be rather strong in practice, notably the joint convexity property. Indeed, there are numerous situations in which the decision variable multiplies the random variable, as in the portfolio problem presented in §3, or in a quite other domain, when one wants to model the breakdown of an actuator, in which case the random variable must be able to kill the action the decision variable. In all those situations, the joint convexity property is not realistic.

2.2 Mathematical Approach for Programming under Probability Constraint

Before explaining our resolution strategy, we review some basic results on the stochastic Arrow-Hurwciz algorithm [1, 5]. First of all, starting with the deterministic constrained optimization problem (1) and assuming that there exists a saddle point of the Lagrangian (2) over $U^{\mathrm{ad}} \times \mathbb{R}^d_+$, the (deterministic) Arrow-Hurwicz algorithm consists in performing successive minimization and maximization steps to search for this saddle point:

$$u^{k+1} = \Pi_{U^{\mathrm{ad}}}\Big(u^k - \varepsilon^k\big(\nabla_u J(u^k) + \nabla_u \Theta(u^k)\lambda^k\big)\Big), \tag{4a}$$

$$\lambda^{k+1} = \Pi_{+}\Big(\lambda^k + \rho^k\big(\Theta(u^{k+1}) - \alpha\big)\Big), \tag{4b}$$

where $\Pi_{U^{\mathrm{ad}}}$ is the projection onto U^{ad} and Π_+ is the projection onto the cone $\mathbb{R}^d_+$.

2.2.1 Stochastic Arrow-Hurwicz Algorithm

The stochastic Arrow-Hurwicz algorithm is typically used to solve a stochastic optimization problem with constraint in expectation:

$$\min_{u\in U^{\rm ad}} \mathbb{E}\big(j(u,\xi)\big) \quad \text{s.t.} \quad \mathbb{E}\big(\theta(u,\xi)\big) \le \alpha\,, \tag{5}$$

where the calculation of expectations is basically difficult if not impossible. The stochastic algorithm overcomes this difficulty by simultaneously approximating the saddle point and the expectations by a Monte-Carlo like technique. It is in fact a combination of the idea of the Monte-Carlo method with the iterative procedure of gradient methods in optimization.

We do assume that a saddle point $(u^\sharp, \lambda^\sharp)$ over $U^{\rm ad}\times\mathbb{R}^d_+$ exists for the Lagrangian associated with problem (5) (hence $u^\sharp$ is a solution of (5)). Observe that this Lagrangian L (2) is equal to the expectation of $\ell(u,\lambda,\cdot) = j(u,\cdot)+\langle\lambda,\theta(u,\cdot)-\alpha\rangle$. We use unbiased estimates of the gradients of L in u and λ obtained with the corresponding gradients of ℓ evaluated at independent drawings ξ^k of ξ supposed to follow the probability law of ξ. More specifically, at stage k of the algorithm, u^k and λ^k being the current estimates of the solution,

1. we draw an independent sample (according to the law $\mathbb{P}$ of ξ), or we observe a new independent sample ξ^{k+1},
2. we compute the stochastic gradients $\nabla_u j(u^k,\xi^{k+1})$ and $\nabla_u\theta(u^k,\xi^{k+1})$,
3. we update u^{k+1} and λ^{k+1} as follows:

$$u^{k+1} = \Pi_{U^{\rm ad}}\Big(u^k - \varepsilon^k\big(\nabla_u j(u^k,\xi^{k+1}) + \nabla_u\theta(u^k,\xi^{k+1})\,\lambda^k\big)\Big)\,, \tag{6a}$$

$$\lambda^{k+1} = \Pi_+\Big(\lambda^k + \rho^k\big(\theta(u^{k+1},\xi^{k+1})-\alpha\big)\Big)\,. \tag{6b}$$

Under essentially measurability and convexity assumptions, assuming that the Lagrangian of the problem admits a saddle point, and with

$$\sum_{k\in\mathbb{N}}\varepsilon^k = +\infty\,, \quad \sum_{k\in\mathbb{N}}(\varepsilon^k)^2 < +\infty \qquad (\text{and the same for } \rho^k)\,,$$

it is shown in [6] that this algorithm converges in the sense that primal $\{u^k\}_{k\in\mathbb{N}}$ and dual $\{\lambda^k\}_{k\in\mathbb{N}}$ sequences are bounded a.s. and that $\{u^k\}_{k\in\mathbb{N}}$ a.s. weakly converges to some solution $u^\sharp$ of (5).

2.2.2 Mathematical Approach: Strategy and Difficulties

From now on, we assume that the critical or risky event is defined by a single (scalar) inequality, that is, θ is $\mathbb{R}$-valued. Let $\mathbb{I}_{\mathbb{R}^+}$ denote the indicator function of the positive half-line. The principle of our resolution strategy is first to replace the probability constraint by a constraint in expectation

$$-P(u) \le -\pi\,, \tag{7}$$

where $P(u) = \mathbb{P}\big(\theta(u,\xi)\le\alpha\big)$ and this probability is evaluated as an expectation:

$$\mathbb{P}\big(\theta(u,\xi)\le\alpha\big) = \mathbb{E}\Big(\mathbb{I}_{\mathbb{R}^+}\big(\alpha-\theta(u,\xi)\big)\Big)\,, \tag{8}$$

then resort to duality, and lastly resort to the stochastic Arrow-Hurwicz algorithm. Observe that w.r.t. the general formulation (1), Θ is now $-P$ and the constraint level α is now $-\pi$.

There are major difficulties with probability constraints.

- First of all, as we recalled in §2.1, convexity is not preserved. Therefore, existence of a saddle point of the Lagrangian is not granted; in this case, we should resort to *augmented* Lagrangian techniques to increase the chance that a saddle point does exist. However, this raises new problems because the nonlinearities involved in the augmented Lagrangian formula cannot be combined straightforwardly with expectation to yield obvious stochastic gradient algorithms. This issue of using augmented instead of ordinary Lagrangians goes beyond the scope of this paper and is not considered here.
- We rather address here another difficulty: to replace a probability constraint by a constraint in expectation, we need to handle the indicator function (see (8)); but this indicator function involves a discontinuity which may, nevertheless, be smoothed by the expectation operation; however, the stochastic Arrow-Hurwicz algorithm is based on the consideration of a *unique* sample drawn at each iteration; obtaining a stochastic gradient is therefore not trivial. As it will be shown later on, we propose two ways of overcoming this difficulty: Approximation by Convolution (AC) and Finite Differences (FD). Both approaches will lead us to consider algorithms such as (6) in which either a smooth approximation $\widehat{\theta}$ of function θ will be used in both equations (6a) and (6b), or an approximation of its gradient will be used in (6a), leading to a stochastic Arrow-Hurwicz algorithm with *biased* stochastic estimates of the Lagrangian gradients.

3 Structural Difficulties of Programming under Probability Constraint

In this section, we focus on two structural difficulties of optimization problem with probability constraints. The first one is related to the non convexity of probability constraint: we show what are the consequences of this non convexity on the stochastic Arrow-Hurwicz algorithm. The second one concerns the behavior of the probability constraint multiplier in some particular cases.

3.1 The Non Convexity of Probability Constraint

Consider the following optimization problem

$$\min_{u\in\mathbb{R}} \frac{1}{2}(u-1)^2 \quad \text{s.t.} \quad \mathbb{P}(u \leq \xi) \geq \pi\,, \tag{9}$$

where ξ is a normal random variable with mean value -2 and standard deviation 0.1.

In order to point out the first difficulty, we study the qualitative behavior of the underlying deterministic problem, namely that in which the probability constraint is expressed with help of the cumulative distribution function F of ξ: indeed, $\mathbb{P}(u \leq \xi) = 1 - F(u)$ and therefore, the constraint in (9) can be replaced by

$$1 - F(u) \geq \pi \tag{10}$$

without of course altering the corresponding Kuhn-Tucker multiplier.

The Lagrangian of problem (9), with constraint written as in (10), is

$$L(u,\lambda) = \frac{1}{2}(u-1)^2 + \lambda\left(F(u) - 1 + \pi\right);$$

and, the Kuhn-Tucker necessary conditions of optimality allow for the calculation of the solution which is, for example with $\pi = 0.7$,

$$u^\sharp = -2.05244 \qquad \text{and} \qquad \lambda^\sharp = 0.877913\,.$$

As expected, $u^\sharp$ takes the maximal possible value to satisfy the constraint, that is, the $(1-\pi)$-th percentile of the distribution: $F(-2.05244) = 0.3$, so the constraint is active, and saturated.

Figure 1 represents the Lagrangian surface in the (u, λ) domain. For $\lambda = 0$, we recognize

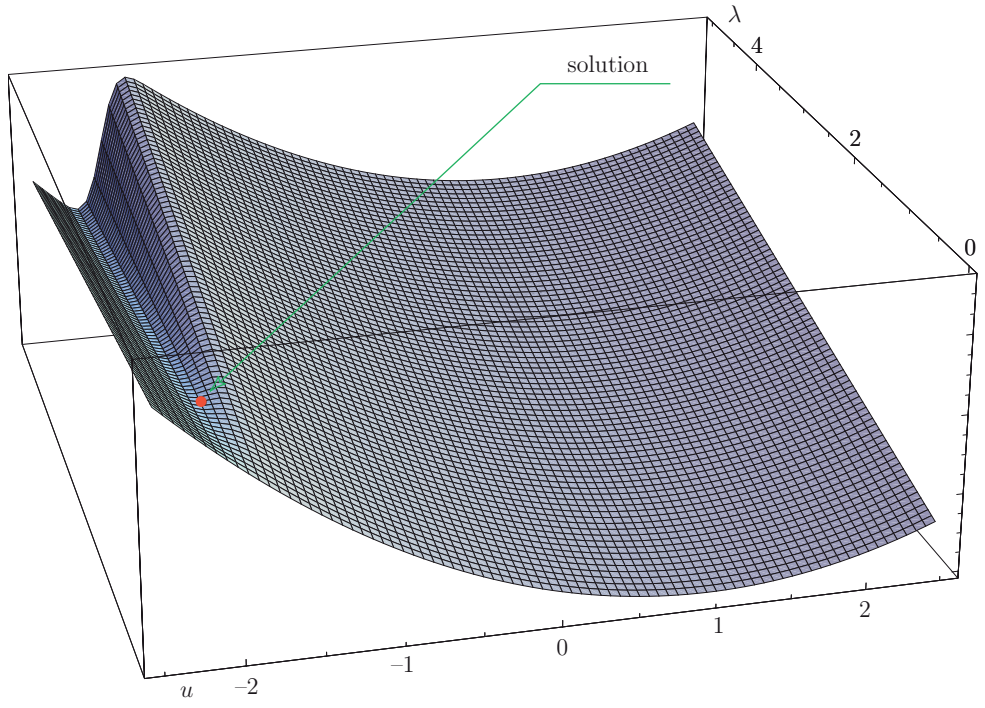

Figure 1: Lagrangian surface

the convex shape of the cost function only. For larger values of λ, the nonconvex form of $F(\cdot)$ shows up more and more, which explains the two valleys.

We insist on the following two points. First, our approach in this paper is based on stochastic estimates of the gradients of the Lagrangian, *not* on their exact computation, which we assume impossible. Naturally, we cannot expect the stochastic algorithm to behave better than its underlying *average* driving vector field, which we will study directly. Second, with some probability distributions there is a way to manipulate the constraints in order to preserve convexity. In particular with normal distributions, the map $\ln(1 - F(\cdot))$ is concave [12], which leads to a convex formulation of the constraint. If we seem to overlook this remark in the following treatment, this is because the difficulty we try to point out in this very simple case is *a fortiori* likely to occur in more general situations when the above clever manipulations are no longer possible: recall that we do not assume knowledge of the distribution of $\theta(u, \xi)$.

Let us now consider the ODE associated with the Arrow-Hurwicz algorithm,

$$\dot{u} = -L'_u(u, \lambda) = -J'(u) - \lambda\, F'(u)\ , \tag{11a}$$

$$\dot{\lambda} = L'_\lambda(u, \lambda) = F(u) - 1 + \pi\ . \tag{11b}$$

At $u^\sharp = 1$, the *unconstrained* optimal solution, one has that $J'(u^\sharp) = 0$ and $F'(u^\sharp) = 1.47 \times 10^{-195}$, because $u^\sharp$ happens to be in the tail of the distribution. Therefore, even for very large values of λ, $L'_u(u^\sharp, \lambda)$ remains very close to 0; in other words, if the (continuous) algorithm (11) is started at (or close to) $(u^\sharp, \lambda)$, for practically any λ, u will stay at $u^\sharp$! At the same time, if $u^\sharp$ doesn't satisfy the constraint, one has that $F(u^\sharp) > 1 - \pi$. It follows that $L'_\lambda(u^\sharp, \lambda) > 0$: λ increases

almost indefinitely! This is better illustrated by the vector field of the ODE, shown in Figure 2.

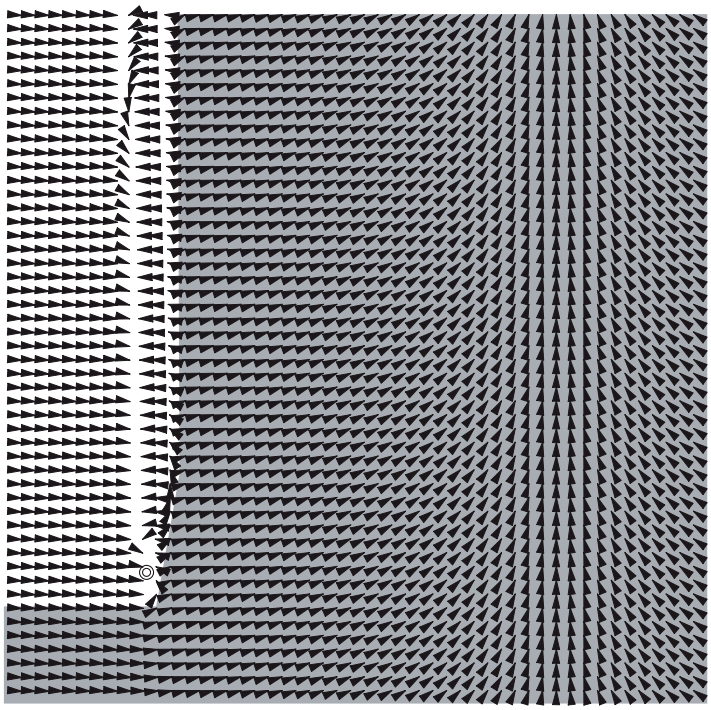

Figure 2: Vector field of the ODE

The white zone corresponds to the basin of attraction of the optimal solution. In the grey zone, the algorithm is driven more or less indefinitely towards large values of λ in a valley corresponding to the unconstrained solution $u^\sharp = 1$.

This example shows that even a deterministic algorithm may, if started on the "wrong" side, wander away from the actual optimal solution. Stochastic versions of the algorithm are expected to behave erratically, and even if the current values of (u^k, λ^k) are in the basin of attraction of the Kuhn-Tucker point, random observations may take the algorithm to other regions away from the optimal solution.

3.2 Degeneracy of the Probability Constraint Multiplier

Let us now consider the following portfolio optimization problem. This very simple problem allows us to point up another structural difficulty of probability constraint. This example will also be used in the remainder of this paper to illustrate our various approaches.

We borrow a capital which we have to pay off at the end of the period with an interest rate l. We can invest a proportion u of this capital at the fixed rate b, invest a proportion v at the random rate ξ, and finally consume the available remainder, which brings a satisfaction measured by a concave nondecreasing function f. We assume of course that $\mathbb{E}(\xi) > l$, in other words, risk is rewarding. We try to maximize the sum of the satisfaction provided by consumption and by the expected final capital. We also want to be in a position to pay off the capital and the interests at the end of the period, with a probability of a least p. In this case, the optimization

problem can be stated as follows:

$$\begin{aligned} &\max_{u,v} \mathbb{E}\big(f(1-u-v)+(1+b)u+(1+\xi)v\big) \\ \text{s.t.}\quad & u \geq 0\,, \quad v \geq 0\,, \quad u+v \leq 1\,, \\ & \mathbb{P}\big((1+b)u+(1+\xi)v \geq 1+l\big) \geq \pi\,. \end{aligned}$$

Let

$$l = 0.15, \qquad b = 0.2, \qquad f(x) = -x^2/2 + 2x\,,$$

$$F(\xi) = \begin{cases} 0 & \text{if} \quad \xi < \bar{\xi} - \sigma\,, \\ \frac{1}{16}\Big(3\big(\frac{\xi-\bar{\xi}}{\sigma}\big)^5 - 10\,\big(\frac{\xi-\bar{\xi}}{\sigma}\big)^3 + 15\,\big(\frac{\xi-\bar{\xi}}{\sigma}\big) + 8\Big) & \text{if} \quad \xi < \bar{\xi} + \sigma\,, \\ 1 & \text{otherwise}\,, \end{cases} \tag{12}$$

where F is the distribution function. For numerical experiments, we set $\bar{\xi} = 0.4$ and $\sigma = 3$. To identify this problem with (3), consider the equivalent minimization problem with cost function

$$j(u,v,\xi) = -f(1-u-v) - (1+b)\,u - (1+\xi)\,v\,.$$

Let also

$$P(u,v) = \mathbb{P}\big((1+b)\,u + (1+\xi)\,v \geq 1+l\big)\,. \tag{13}$$

This problem is now formulated as

$$\min_{u\geq 0, v\geq 0} \mathbb{E}\big(j(u,v,\xi)\big) \quad \text{s.t.} \quad u+v \leq 1, \quad -P(u,v) \leq -\pi$$

with Lagrangian

$$L(u,v;\lambda_1,\lambda_2) = \mathbb{E}\big(j(u,v,\xi)\big) + \lambda_1(u+v-1) + \lambda_2(\pi - P(u,v))\,.$$

Figure 3 represents the *optimal cost* as a function of probability level π. We observe that

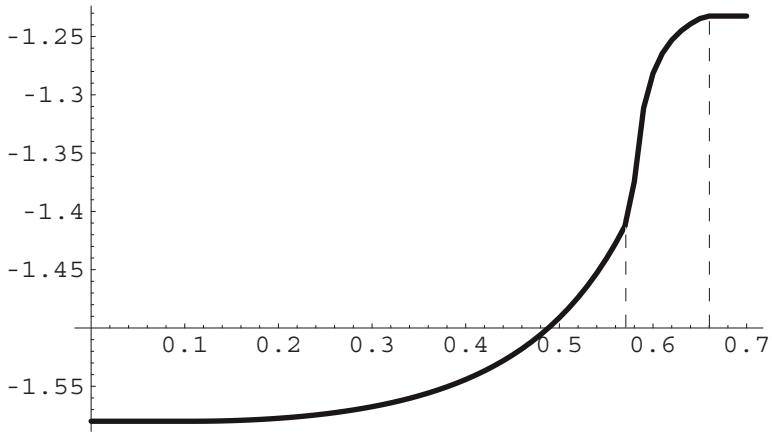

Figure 3: Optimal cost

this function is not convex. In fact, it is convex for probability levels below 0.57. For probability levels close to 0.57, the risk of not being in a position to pay off the capital and the interests is important; the investment in the risky asset v decreases to zero, whereas simultaneously, that

in the secure asset u increases. The optimal cost, which was until then a convex function of the required probability level, becomes concave. Above 0.65, v is zero, u is equal to $(1+l)/(1+b) = 0.95833$ in order to satisfy the probability constraint, and the optimal cost becomes constant.

This example shows another structural difficulty of optimization under a probability constraint, namely the degeneracy of the probability constraint multiplier. Indeed, for π small enough, the secure asset, u is zero at optimum. For π large enough, the risky asset, v, is zero at optimum. In the latter case, the event $\{(1+b)\,u + (1+\xi)v \geq 1+l\}$ can only have probability 0 or 1. That is to say, at $v = 0$ and $u = (1+l)/(1+b) = 0.95833$, the function (13) exhibits a discontinuity (see Figure 4).

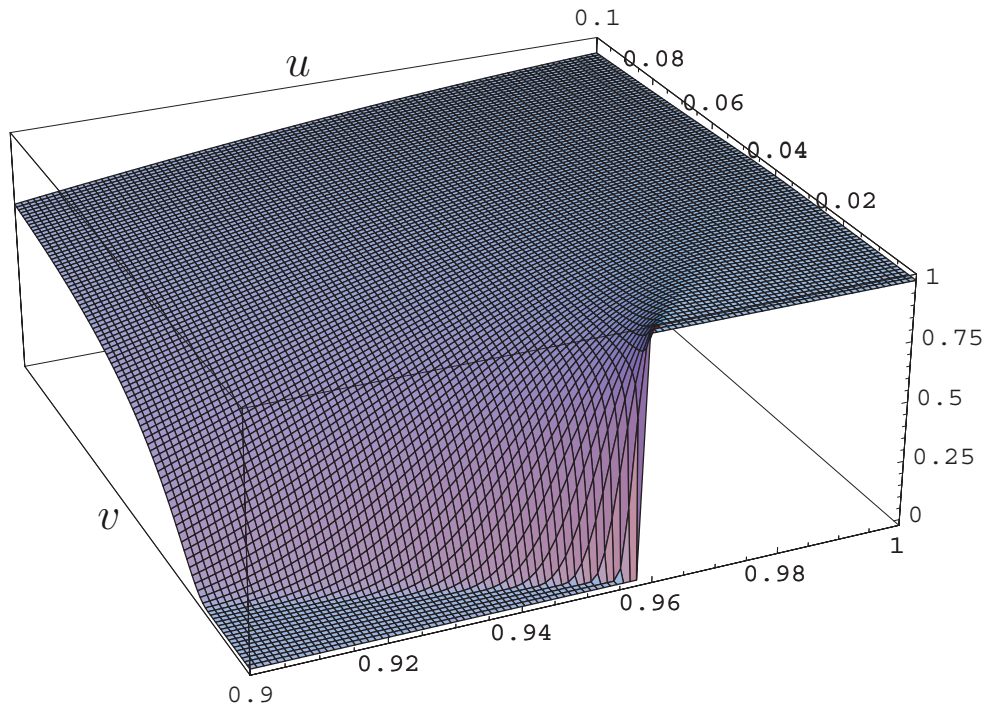

Figure 4: Graph of the probability function for $(u, v) \in [0.9, 1] \times [0, 0.1]$

When π is large and $v^\sharp = 0$, the probability (13) can only take values 0 or 1 (depending on the value of u), that is, this probability is strictly larger than the required level π when the constraint is met. Clearly the constraint is not "saturated", because there is no equality, and consequently the corresponding multiplier is zero; small changes in π will not affect the solution. However, the constraint is "active", that is, the solution $(u^\sharp, v^\sharp)$ of the problem *with* the probability constraint is different from the solution (u^*, v^*) *without* it.

In the remainder, in order to guarantee existence of a saddle point of the Lagrangian, we consider only probability levels below 0.57 (otherwise, one should resort to augmented Lagrangian techniques, but, as it has been said earlier, this issue goes beyond the scope of this paper). For example, for a probability level of 0.24, the primal-dual optimal solution is

$$u^\sharp = 0\,, \quad v^\sharp = 0.50407\,, \quad \lambda_1^\sharp = 0\,, \quad \lambda_2^\sharp = 0.08815\,. \tag{14}$$

4 Stochastic Estimates of Probability Function Gradient

As it was mentioned at §2.2.2, in order to use a stochastic Arrow-Hurwicz algorithm, we need to handle the probability function gradient, that is, to obtain a stochastic estimate of the gradient of (see (8))

$$P(u) = \mathbb{E}\Big(\mathbb{I}_{\mathbb{R}^+}\big(\alpha - \theta(u,\xi)\big)\Big)\,. \tag{15}$$

It is well known that this gradient is difficult to compute; we may refer to [14] for a discussion of this topic. Recall that in our case, replacing the probability constraint by a constraint in expectation raises the difficulty of handling an indicator function, which is a discontinuous function. One way of dealing with this problem is to appeal to a technique based on convolution to derive a smooth approximation of this discontinuous function. Alternatively, we can obtain a stochastic estimate of the gradient of this function, based on a single sample drawing of ξ, by appealing to a finite difference technique, and we rely upon the multiplication of such drawings along the iterative algorithm to smooth out this crude estimate.

4.1 Approximation by Convolution Method (AC)

4.1.1 General Theory

The basic principle of this approach is to smooth out the indicator function appearing in (15) so that differentiation underneath expectation becomes possible. Consider a smooth function $h : \mathbb{R} \to \mathbb{R}$ with the following properties : h as a unique maximum at $x = 0$,

$$\forall x, \quad h(x) \geq 0; \quad h(x) = h(-x); \qquad \int_{-\infty}^{+\infty} h(x)\,dx = 1\,. \tag{16}$$

We will give a few examples of such functions later on and will consider only functions with finite support although this is not absolutely necessary. With any other function $\phi : \mathbb{R} \to \mathbb{R}$, and r a small positive number, the convolution

$$\phi_r(x) = \frac{1}{r}\int_{-\infty}^{\infty} \phi(y) h\Big(\frac{x-y}{r}\Big)\,dy\,,$$

can be viewed as an approximation of ϕ since $h(\cdot/r)/r$ approximates the Dirac function (in the sense of convergence of distributions) at 0 when r tends to zero. The function ϕ_r is differentiable with

$$\phi_r'(x) = \frac{1}{r^2}\int_{-\infty}^{\infty} \phi(y)\, h'\Big(\frac{x-y}{r}\Big)\,dy\,.$$

This technique is widely known as the "mollifier" technique [7]. We now apply it to $\mathbb{I}_{\mathbb{R}^+}$: recall (15) and define

$$\begin{aligned} P_r(u) &= \frac{1}{r}\mathbb{E}\left(\int_{-\infty}^{+\infty} \mathbb{I}_{\mathbb{R}^+}(y)\, h\Big(\frac{\alpha - \theta(u,\xi) - y}{r}\Big)\,dy\right) \\ &= \frac{1}{r}\mathbb{E}\left(\int_{0}^{+\infty} h\Big(\frac{y - \alpha + \theta(u,\xi)}{r}\Big)\,dy\right) \end{aligned}$$

(here, we have used the fact that h is an even function)

$$= \mathbb{E}\big(p_r(u,\xi)\big) \tag{17}$$

with

$$p_r(u,\xi) = \frac{1}{r}\int_0^{+\infty} h\Big(\frac{y-\alpha+\theta(u,\xi)}{r}\Big)\,dy\,. \tag{18}$$

Then

$$\begin{aligned}(p_r)'_u(u,\xi) &= \frac{1}{r^2}\,\theta'_u(u,\xi)\int_0^{+\infty} h'\Big(\frac{y-\alpha+\theta(u,\xi)}{r}\Big)\,dy\\ &= -\frac{1}{r}\,h\Big(\frac{\theta(u,\xi)-\alpha}{r}\Big)\,\theta'_u(u,\xi)\,,\end{aligned} \tag{19}$$

and clearly

$$P'_r(u) = \mathbb{E}\big((p_r)'_u(u,\xi)\big)\,. \tag{20}$$

Therefore, for any sample ξ, (19) can be considered as a stochastic estimate of $P'(u)$, albeit a *biased* one; however, this bias vanishes when r approaches 0. In what follows, we evaluate the bias and the variance of this estimate as a function of r.

Remark 1. In the same way, according to (17), (18) can be considered a biased estimate of $P(u)$ whereas

$$p(u,\xi) = \mathbb{I}_{\mathbb{R}^+}\big(\alpha-\theta(u,\xi)\big) \tag{21}$$

is an unbiased one. In Equation (6b) of the iterative algorithm, we may either use the unbiased estimate or the biased one, consistently with that used in (6a) for Θ'. The latter option has the advantage of preserving the specific geometry of vector fields of Arrow-Hurwicz algorithms (with some symmetry, or skew-symmetry, properties, according to the point of view). The former option may seem preferable as long as it avoids seemingly unnecessary bias or approximation. Both options will be tested later on in §6. Therefore, the next theorem deals with both the estimates (18) and (19) in order to cover all variants.

Theorem 2. *The random variable (or vector) ξ is supposed to admit a density $q(\xi)$. For the random variable $X_u(\cdot) = \theta(u,\cdot)$ depending on the parameter u, we assume that the induced probability law also admits a density denoted $q_{X_u}(x)$ and that this density is at least twice continuously differentiable with L^1 first and second order derivatives. Then, for any sample drawing ξ following the probability density q, the expression* (18) *provides a biased estimate of $P(u)$ with a bias in $\mathrm{O}(r^2)$ and a variance in $\mathrm{O}(1)$.*

For the pair of random variables (or vectors) $\big(X_u(\cdot),Y_u(\cdot)\big) = \big(\theta(u,\cdot),\theta'_u(u,\cdot)\big)$ depending on the parameter u, we assume that the induced joint probability law admits a density denoted $q_{X_uY_u}(x,y)$ and that this density is at least twice continuously differentiable in x with integrable L^1 first and second order derivatives. Then, for any sample drawing ξ following the probability density q, the expression (19) *provides a biased estimate of $P'(u)$ with a bias in $\mathrm{O}(r^2)$ and a variance in $\mathrm{O}(1/r)$.*

Proof. Consider first (17)–(18). With the induced probability law for the random variable X_u, one has that

$$P_r(u) = \frac{1}{r}\int_{-\infty}^{+\infty}\int_0^{+\infty} h\Big(\frac{y-\alpha+x}{r}\Big)\,q_{X_u}(x)\,dy\,dx\,.$$

Using Fubini theorem and the change of variable $z=(y-\alpha+x)/r$ in the integral in x yields

$$P_r(u) = \int_0^{+\infty}\int_{-\infty}^{+\infty} h(z)\,q_{X_u}(rz-y+\alpha)\,dz\,dy\,.$$

With the smoothness assumptions on q_{X_u}, the Taylor expansion of this term for r near 0 yields

$$\begin{aligned} P_r(u) &= \int_0^{+\infty} \int_{-\infty}^{+\infty} h(z) \left(q_{X_u}(\alpha - y) + rz\, q'_{X_u}(\alpha - y) + \frac{r^2 z^2}{2}\, q''_{X_u}(\alpha - y) + \mathrm{O}(r^3) z^3 \right) dz\, dy \\ &= \int_0^{+\infty} q_{X_u}(\alpha - y)\, dy + \frac{r^2}{2} \sigma_h^2 \int_0^{+\infty} q''_{X_u}(\alpha - y)\, dy + \mathrm{O}(r^3) \end{aligned}$$

by using (16) on the one hand and by introducing

$$\sigma_h^2 = \int_{-\infty}^{+\infty} z^2\, h(z)\, dz \tag{22}$$

on the other hand. The term of order 0 in r can be written as

$$\int_{-\infty}^{\alpha} q_{X_u}(t)\, dt$$

and, as such, is recognized to be equal to $\mathbb{P}(X_u \leq \alpha)$, that is, $P(u)$. Therefore, $P_r(u)$ differs from $P(u)$ by an $\mathrm{O}(r^2)$ term (proportional to σ_h^2).

As for the variance of the estimate (18), it is equal to the second order moment $\mathbb{E}\Big(\big(p_r(u,\xi)\big)^2 \Big)$ from which the square of $\mathbb{E}\big(p_r(u,\xi)\big)$ must be subtracted. The latter is close to $\big(P(u)^2\big)$ up to a term of order $\mathrm{O}(r^2)$. Therefore we concentrate on the second order moment which can be written, according to (18),

$$\begin{aligned} \mathbb{E}\Big(\big(p_r(u,\xi)\big)^2 \Big) &= \frac{1}{r^2} \mathbb{E}\left(\Big(\int_0^{+\infty} h\Big(\frac{y - \alpha + \theta(u,\xi)}{r} \Big)\, dy \Big)^2 \right) \\ &= \frac{1}{r^2} \int_{-\infty}^{+\infty} \Big(\int_0^{+\infty} h\Big(\frac{y - \alpha + x}{r} \Big)\, dy \Big)^2 q_{X_u}(x)\, dx \\ &= \int_{-\infty}^{+\infty} \Big(\int_{\frac{x-\alpha}{r}}^{+\infty} h(z)\, dz \Big)^2 q_{X_u}(x)\, dx \end{aligned}$$

using the change of variable $z = (y - \alpha + x)/r$ in the integral in y,

$$\leq \int_{-\infty}^{+\infty} \Big(\int_{-\infty}^{+\infty} h(z)\, dz \Big)^2 q_{X_u}(x)\, dx$$

since $h(\cdot) \geq 0$,

$$= 1$$

according to (16) (last equality) and the fact that q_{X_u} is a probability density.

The proof regarding the bias of $P'_r(u)$ w.r.t. $P'(u)$ may follow one of the following two paths: either a similar result is proved for the derivative of a function whenever the function itself is approximated by another function up to an $\mathrm{O}(r^2)$ term; or, with (19)–(20), we perform similar calculations to those we have just performed with (17)–(18). Let us sketch this second path. Considering (19)–(20) and the pair $\big(X_u(\cdot), Y_u(\cdot)\big)$, we have that

$$P'_r(u) = -\frac{1}{r} \int \int h\Big(\frac{x - \alpha}{r} \Big)\, y\, q_{X_u Y_u}(x,y)\, dx\, dy$$

(remember y may be a vector of the same dimension as u and dy should be understood adequately). From here, we proceed as previously with the change of variable $z = (x - \alpha)/r$ which yields

$$P'_r(u) = -\int\int h(z)\, y\, q_{X_u Y_u}(rz + \alpha, y)\, dz\, dy\,.$$

Then, a Taylor expansion of $q_{X_u Y_u}$ w.r.t. its first argument for r near 0 yields, for the same reasons as previously,

$$P'_r(u) = -\int y\, q_{X_u Y_u}(\alpha, y)\, dy + \frac{r^2}{2}\sigma_h^2 \int \frac{\partial^2 q_{X_u Y_u}(\alpha, y)}{\partial x^2}\, y\, dy + \mathrm{O}(r^3)\,. \tag{23}$$

Assuming that the first term in the right-hand side above is equal to $P'(u)$ (see Claim 3 hereafter), we obtain again that $P'_r(u)$ differs by an $\mathrm{O}(r^2)$ term.

The variance is equal to the second order moment $\mathbb{E}\Big(\big(p'_r(u,\xi)\big)^2\Big)$ from which we must subtract $\Big(\mathbb{E}\big(p'_r(u,\xi)\big)\Big)^2$. The latter is close to $\big(P(u)\big)^2$ up to $\mathrm{O}(r^2)$. As for the former, we have that

$$\begin{aligned}
\mathbb{E}\Big(\big(p'_r(u,\xi)\big)^2\Big) &= \frac{1}{r^2}\int h^2\Big(\frac{\theta(u,\xi) - \alpha}{r}\Big)\big(\theta'_u(u,\xi)\big)^2 q(\xi)\, d\xi \\
&= \frac{1}{r^2}\int\int h^2\Big(\frac{x - \alpha}{r}\Big)\, y^2 q_{X_u Y_u}(x, y)\, dx\, dy \\
&= \frac{1}{r}\int\int h^2(z)\, y^2 q_{X_u Y_u}(rz + \alpha, y)\, dz\, dy\,.
\end{aligned}$$

From here, we proceed as earlier with a Taylor expansion for r close to 0, and it should be clear that the above expression is of order $1/r$ with a coefficient which can be bounded by a term proportional to the square of the L^2 norm of h. The same consideration is still valid for the variance itself. □

Claim 3. The first term in the right-hand side of (23) is equal to $P'(u)$. We sketch the proof of this fact here. For any smooth function $f : \mathbb{R} \to \mathbb{R}$, consider

$$F(u) = \mathbb{E}\Big(f\big(\theta(u,\xi)\big)\Big) = \int f\big(\theta(u,\xi)\big)\, q(\xi)\, d\xi = \int f(x)\, q_{X_u}(x)\, dx\,.$$

Then,

$$F'(u) = \int f'\big(\theta(u,\xi)\big)\, \theta'_u(u,\xi)\, q(\xi)\, d\xi = \int\int f'(x)\, y\, q_{X_u Y_u}(x, y)\, dx\, dy\,.$$

Integrating by parts in the integral in x, one gets

$$F'(u) = -\int\int f(x)\, y\, \frac{\partial q_{X_u Y_u}}{\partial x}(x, y)\, dx\, dy\,.$$

If f is not smooth enough for this calculation to be immediately justified, one can consider a sequence of smooth approximations converging to f in order to establish this formula. Let us now use it for $f(\cdot) = \mathbb{I}_{\mathbb{R}^+}(\alpha - \cdot)$. Then $F(u) = P(u)$ and

$$\begin{aligned}
P'(u) &= -\int y \int \mathbb{I}_{\mathbb{R}^+}(\alpha - x)\, \frac{\partial q_{X_u Y_u}}{\partial x}(x, y)\, dx\, dy \\
&= -\int y \int_{-\infty}^{\alpha} \frac{\partial q_{X_u Y_u}}{\partial x}(x, y)\, dx\, dy \\
&= -\int y\, q_{X_u Y_u}(\alpha, y)\, dy\,,
\end{aligned}$$

which is exactly the expected result.

Remark 4. As a side remark, observe that

$$q_{X_u}(\alpha) = \int q_{X_u Y_u}(\alpha, y)\, dy\,,$$

and that

$$q_{X_u Y_u}(\alpha, y)/q_{X_u}(\alpha) = q_{Y_u}(y \mid X_u = \alpha)\,,$$

that is, the conditional density of Y_u knowing that $X_u = \alpha$. Therefore, the first term in the right-hand side of (23) can be written as $-\mathbb{E}\big(Y_u \mid X_u = \alpha\big) \times q_{X_u}(\alpha)$. We conclude that

$$P'(u) = -q_{\theta(u,\cdot)}(\alpha) \times \mathbb{E}\big(\theta'_u(u,\cdot) \mid \theta(u,\cdot) = \alpha\big)\,.$$

Remark 5. Observe that, although we started with the idea of a smooth function h, the expression (19) of the estimate and the analysis in the proof of Theorem 2 does not involve more than the function h itself (not its derivatives), so that we may as well consider non smooth functions (and even discontinuous functions at the ends of its support).

In conclusion, the variance of the stochastic estimate (19) blows up like A/r as r goes to 0 (where A can be bounded from above by something proportional to the square of the L^2 norm of h), that of (18) remains of order O(1), whereas the square of the bias of both estimates goes to 0 as $B^2 r^4$ (where B is proportional to σ_h^2 — see (22)). If the estimate of $P'(u)$ is rather based on the average of N expressions as (19) for N independent drawings of ξ, the variance will blow up as $A/(Nr)$ whereas the square of the bias will still behave as $B^2 r^4$. Therefore, the best trade-off between variance and bias is realized by that r which minimizes the mean square error (MQE; this is the sum of the variance and of the square of the bias) equal to $A/(Nr) + B^2 r^4$: the "best" r is thus $\big(A/(4B^2N)\big)^{1/5}$. This yields a MQE estimated to $(5A^{4/5}B^{2/5})/(4N)^{4/5}$. Therefore, in the choice of function h, it is meaningful to pay attention to the quantity $\sigma_h^{4/5}\|h\|_{L^2}^{8/5}$. Remember B is proportional to σ_h^2 and A is proportional to $\|h\|_{L^2}^2$.

The bias of the AC estimate goes to 0 with r: this parameter r allows for a trade-off between mean and variance which should be adapted to the number of samples available (as just discussed) or visited in one run in the context of an iterative algorithm, as discussed later on in §5.3.

4.1.2 Practical Aspects and Application to Example of §3.2

Define

$$I(x) = \begin{cases} 1 & \text{if } -1 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Table 1 proposes 6 functions with bounded supports that can play the role of function h (see (16)) and compares them from the point of view of their constants $\sigma_h^{4/5}\|h\|_{L^2}^{8/5}$ (last column), the relevance of which has just been explained. The column $h(0)$ is provided to help identifying the functions with their graphs displayed in Figure 5.

We observe that the fourth function, namely $h(x) = 3(1-x^2)I(x)/4$ is the one to retain because it offers the smallest value in the last column of the table. We now apply the technique to the example of §3.2 again. The estimates for $P'_u(u,v)$ and $P'_v(u,v)$ based on this technique and on a given sample ξ read as follows:

$$(p_r)'_u(u,v,\xi) = \frac{1+b}{r}\, h\Big(\frac{(1+\alpha) - (1+b)u - (1+\xi)v}{r}\Big)\,; \tag{24a}$$

$$(p_r)'_v(u,v,\xi) = \frac{1+\xi}{r}\, h\Big(\frac{(1+\alpha) - (1+b)u - (1+\xi)v}{r}\Big)\,. \tag{24b}$$

h	$h(0)$	σ_h^2	$\|h\|_{L^2}^2$	$\sigma_h^{4/5}\|h\|_{L^2}^{8/5}$
$I(x)$	0.5000	0.3333	0.5000	0.3701
$(1-\|x\|)I(x)$	1.000	0.1667	0.6667	0.3531
$\pi\cos(\pi x/2)I(x)/4$	0.7854	0.1894	0.6169	0.3492
$3(1-x^2)I(x)/4$	0.7500	0.2000	0.6000	0.3491
$15(1-x^2)^2I(x)/16$	0.9375	0.1429	0.7143	0.3508
$35(1-x^2)^3I(x)/32$	1.0938	0.1111	0.8159	0.3529

Table 1: Various h functions

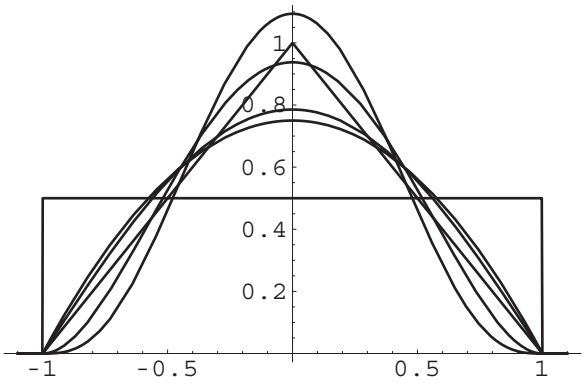

Figure 5: Several possible h functions

This MQE will be compared with that obtained by the next approach, namely finite differences.

We performed some exact computations of bias and variance with the help of *Mathematica* for those estimates evaluated at the optimal solution (see (14)) and with h equal to the fourth function in Table 1. We have found:

$$\begin{aligned}
\mathbb{E}(p_r)'_u(u^\sharp, v^\sharp, \cdot) &= 0.62 - 0.096r^2 + 0.012r^4\,,\\
\mathbb{V}\mathrm{ar}(p_r)'_u(u^\sharp, v^\sharp, \cdot) &= 0.45/r - 0.39 - 0.05r + \mathrm{O}(r^2)\,,\\
\mathbb{E}(p_r)'_v(u^\sharp, v^\sharp, \cdot) &= 1.18 - 0.36r^2 + 0.06r^4\,,\\
\mathbb{V}\mathrm{ar}(p_r)'_v(u^\sharp, v^\sharp, \cdot) &= 1.62/r - 1.39 - 0.35r + \mathrm{O}(r^2)\,.
\end{aligned}$$

If the estimates are based on the average over N samples, the MQE of the AC estimates are obtained by considering $\mathbb{V}\mathrm{ar}(r)/N + (\mathbb{E}(r))^2 - (\mathbb{E}(0))^2$. In those expressions, we consider the terms in $1/Nr$ and r^4 *only* in order to tune r as a function of N. This computation is done for the sum of the MQE's related to the two components of p'_r (that is, for the mean square norm of the vector estimate error — we denote it $\mathrm{MQE}(r, N)$). This yields $r = 1.30\,N^{-1/5}$. Finally, we plug this value of r into $\mathrm{MQE}(r, N)$ to get the following function of N (again, calculations are exact, up to *Mathematica* accuracy, even if results are displayed in a truncated form):

$$\frac{1.98}{N^{4/5}} - \frac{1.78}{N} + \mathrm{O}(N^{-6/5})\,. \tag{25}$$

4.2 Finite Differences (FD)

4.2.1 General Theory

The idea here is simply to evaluate the derivative w.r.t. each component u_j of the expression inside expectation in (15) by the variation of this quantity, caused by, and divided by, the symmetric variation $(u_j+c)-(u_j-c)=2c$ for a sample ξ. We denote $\mathbf{1}_j$ the vector of the same dimension as u with a 1 in the j-th component and 0 elsewhere. The FD stochastic estimate of P'_{u_j} is

$$\widetilde{\nabla^c_{u_j}p}(u,\xi)=\frac{\mathbb{I}_{\mathbb{R}^+}\big(\alpha-\theta(u+c\mathbf{1}_j,\xi)\big)-\mathbb{I}_{\mathbb{R}^+}\big(\alpha-\theta(u-c\mathbf{1}_j,\xi)\big)}{2c}\,. \tag{26}$$

It is wise to use the same sample ξ for the evaluation at $u+c$ and $u-c$ in order to reduce variance. A symmetric difference around u is also recommended. We notice that [11] includes FD under a.s. continuity assumptions, which is not our case here, because the indicator functions are discontinuous.

The following theorem provides the analysis of bias and variance of this estimate w.r.t. the parameter c.

Theorem 6. *If P (see (15)) is three times continuously differentiable with bounded derivatives, the expression (26) provides a biased estimate of P'_{u_j} with a bias in $\mathrm{O}(c^2)$. If*

(H1) *$\theta(\cdot,\xi)$ is differentiable with derivatives bounded uniformly in ξ;*

(H2) *the probability measure of ξ has a density;*

(H3) *$\theta(u,\cdot)$ is twice differentiable and, for all u, and for every solution $\hat\xi$ of $\theta(u,\xi)=\alpha$, we have that $\theta'_\xi(u,\hat\xi)\neq 0$;*

then the variance of estimate (26) is in $\mathrm{O}(c^{-1})$.

If **(H1)** *and* **(H2)** *still hold true but* **(H3)** *is replaced by*

(H4) *$\theta(u,\cdot)$ is three times differentiable and, whenever $\theta(u,\hat\xi)=\alpha$ for some $\hat\xi$, and $\theta'_\xi(u,\hat\xi)=0$, we have that $\theta''_{\xi^2}(u,\hat\xi)\neq 0$;*

then the variance of estimate (26) is in $\mathrm{O}(c^{-3/2})$.

Finally, under no particular assumptions on g, the best bound for the variance is in $\mathrm{O}(c^{-2})$.

Proof. With the smoothness assumption on P, one has that

$$\begin{aligned}\mathbb{E}\widetilde{\nabla^c_{u_j}p}(u,\cdot)-P'_{u_j}(u)&=\frac{P(u+c\mathbf{1}_j)-P(u-c\mathbf{1}_j)-2c\,P'_{u_j}(u)}{2c}\\&=\frac{c^2}{6}P'''_{u_j^3}(u)+\mathrm{O}(c^3)\,,\end{aligned}$$

which proves the claim on the bias.

To evaluate the variance of (26), we study its second order moment which differs from the variance by $\big(P'_{u_j}(u)\big)^2$ up to terms in $\mathrm{O}(c^2)$ as we have just seen. Consider

$$\begin{aligned}\mathbb{E}\Big(\widetilde{\nabla^c_{u_j}p}(u,\xi)\Big)^2&=\mathbb{E}\left(\frac{\mathbb{I}_{\mathbb{R}^+}\big(\alpha-\theta(u+c\mathbf{1}_j,\xi)\big)-\mathbb{I}_{\mathbb{R}^+}\big(\alpha-\theta(u-c\mathbf{1}_j,\xi)\big)}{2c}\right)^2\\&=\frac{1}{4c^2}\Big(\mathbb{P}\big(\{\theta(u+c\mathbf{1}_j,\xi)\le\alpha\}\cap\{\theta(u-c\mathbf{1}_j,\xi)>\alpha\}\big)\\&\qquad+\mathbb{P}\big(\{\theta(u+c\mathbf{1}_j,\xi)>\alpha\}\cap\{\theta(u-c\mathbf{1}_j,\xi)\le\alpha\}\big)\Big)\,,\end{aligned}$$

those two events being of course disjoint.

Using the mean value theorem (or Taylor representation) for the function $\theta(\cdot,\xi) \in C^1$, we have that $\theta(u + c\mathbf{1}_j, \xi) = \theta(u,\xi) + c\theta'_{u_j}\big(v^+(\xi),\xi\big)$, and similarly $\theta(u - c\mathbf{1}_j, \xi) = \theta(u,\xi) - c\theta'_{u_j}\big(v^-(\xi),\xi\big)$. Therefore,

$$\mathbb{E}\Big(\widetilde{\nabla^c_{u_j} p}(u,\xi)\Big)^2 = \frac{1}{4c^2}\Bigg(\mathbb{P}\Big(\big\{\alpha + c\theta'_{u_j}\big(v^-(\xi),\xi\big) < \theta(u,\xi) \le \alpha - c\theta'_{u_j}\big(v^+(\xi),\xi\big)\big\}\Big) + \mathbb{P}\Big(\big\{\alpha - c\theta'_{u_j}\big(v^+(\xi),\xi\big) < \theta(u,\xi) \le \alpha + c\theta'_{u_j}\big(v^-(\xi),\xi\big)\big\}\Big)\Bigg), \quad (27)$$

Thanks to **(H1)**, we can bound each of the above two probabilities by

$$\mathbb{P}\big(\{\theta(u,\xi) \in (\alpha - Kc, \alpha + Kc]\}\big), \quad (28)$$

where K is the uniform bound on θ'_{u_j}.

Our goal is now to evaluate the behavior of this probability when c is approaching 0. Let m be the dimension of ξ; g is supposed to be $\mathbb{R}$-valued. Consider any solution $\hat{\xi}$ of

$$\theta(u,\xi) = \alpha\,. \quad (29)$$

The case when no such solution exists for some u will be discussed later on at Remark 7. If $\theta'_\xi(u,\hat{\xi}) \neq 0$ as assumed in **(H3)**, then the manifold of solutions of (29) is locally of dimension less than or equal to $m-1$. The set of ξ's involved in the event in (28) is locally a set with a "backbone" given by this manifold around $\hat{\xi}$, and a "thickness" which is proved to be of order $\mathrm{O}(c)$. Indeed, with a Taylor expansion of $\theta(u,\cdot)$ around $\hat{\xi}$, we get

$$\theta(u,\hat{\xi}+y) = \alpha + \Big\langle \theta'_\xi(u,\hat{\xi}), y\Big\rangle + \mathrm{O}(\|y\|^2)\,.$$

In this expression, we need only consider y's which are (asymptotically as c goes to 0) parallel to the gradient $\theta'_\xi(u,\hat{\xi})$ (that is, the component in the kernel of the linear form defined by this gradient is useless). It should now be obvious that to match variations of g around α which are of order c, we need only consider y's which are also of order c in norm. If this holds true for any $\hat{\xi}$ in the manifold of solutions of (29), then the probability (28) is of order $\mathrm{O}(c)$ and the second order moment (27) of our estimate (and consequently the variance too) is bounded by an $\mathrm{O}(c^{-1})$.

If **(H3)** does not hold but **(H4)** does, then the same reasoning can be repeated (for a Taylor expansion of the next order) with y's which are now orthogonal to the kernel of the Hessian $\theta''_{\xi^2}(u,\hat{\xi})$ (this component is non zero thanks to **(H4)**) and it should be clear that to compensate for variations of g of order c, we now need y's which are of order $\mathrm{O}(c^{1/2})$ in norm. This also gives the order of the probability (28) and then, the bound on the variance is in $\mathrm{O}(c^{-3/2})$.

We could continue like that by removing assumption **(H4)** but introducing an assumption **(H5)**, and so on and so forth. Ultimately, with no particular assumptions, (28) is of order $\mathrm{O}(1)$ and the variance is of order $\mathrm{O}(c^{-2})$. □

Remark 7. Suppose that for some u, there exists no solutions to (29). Then, since g is assumed to be at least continuous in ξ, this means that for all ξ, $\theta(u,\xi)$ is always either strictly less or strictly greater than α, in which cases $P(u)$ (see (15)) assumes either the value 1 or 0 (which are extreme values for P).

If $\theta(u,\cdot)$ can be bounded away from α, then the probability (28) will be 0 for c small enough. This is the good case for the variance of the estimate. But $\theta(u,\cdot)$ may also approach α asymptotically, and, with heavy tails for the density q of ξ, it is not possible to give a better bound for (28) than O(1). Here is an example. Let $\theta(u,\xi) = u - e^{-\xi}$ (u and ξ are both scalar). Consider the probability $\mathbb{P}(\theta(0,\xi)) \in [-c,c]$ for c small, that is, $\mathbb{P}(\xi \geq -\ln c)$. Assume the density $q(\xi)$ has a positive support and that it is equal to $a(1+\xi)^{-(1+a)}\,\mathbb{I}_{\mathbb{R}^+}(\xi)$ with a an arbitrary small positive number. Then $\mathbb{P}(\xi \geq -\ln c) = (1-\ln c)^{-a}$. For c positive and below e, $(1-\ln c)^{-1} \geq c$, hence this probability is larger than c^a. Since a is positive and arbitrarily small, we cannot clearly make this case enter the case of better bounds obtained with assumptions **(H3)** or **(H4)**.

4.2.2 Application to the Example and Comparison with the AC Method

We have used (26) (for the two components of the gradient, that in u and that in v) to our example and evaluated, once again with the help of *Mathematica*, the mean and variance of those estimates at the optimal solution (14). The results are as follows:

$$\begin{aligned}
\mathbb{E}\widetilde{\nabla^c_u p}(u^\sharp, v^\sharp, \cdot) &= 0.62 - 0.23c^2 + 0.06c^4 + \mathrm{O}(c^7)\,,\\
\mathbb{V}\mathrm{ar}\widetilde{\nabla^c_u p}(u^\sharp, v^\sharp, \cdot) &= \frac{0.31}{c} - 0.39 - 0.12c + \mathrm{O}(c^2)\,,\\
\mathbb{E}\widetilde{\nabla^c_v p}(u^\sharp, v^\sharp, \cdot) &= 1.18 - 1.49c^2 - 42.25c^4 - 199.41c^6 + \mathrm{O}(c^7)\,,\\
\mathbb{V}\mathrm{ar}\widetilde{\nabla^c_v p}(u^\sharp, v^\sharp, \cdot) &= \frac{0.59}{c} - 0.39 - 0.74c + \mathrm{O}(c^2)\,.
\end{aligned}$$

Following the same procedure as for the AC estimate, the MQE for the gradient vector estimate based on N independent samples is obtained by $\mathbb{V}\mathrm{ar}(c)/N + (\mathbb{E}(c))^2 - (\mathbb{E}(0))^2$; in this expression, the dominant terms in $1/Nc$ and in c^4 only are retained to tune c as a function of N. This yields $c = 0.63N^{-1/5}$ and an optimal MQE equal to:

$$\frac{1.79}{N^{4/5}} - \frac{1.78}{N} + \mathrm{O}(N^{-6/5})\,. \tag{30}$$

Compared with (25) which was obtained with the AC estimate, this is asymptotically slightly better. However, a more careful inspection with complete expressions of the MQEs shows that this conclusion becomes true only for N above about 11000. Hence one may say that the AC and the FD methods yield approximately the same performances.

5 Convergence Analysis

5.1 Stochastic Algorithms

Consider algorithm (6). With $\Theta(u) = \mathbb{E}\big(\theta(u,\xi)\big)$ and $J(u) = \mathbb{E}j(u,\xi)$, an equilibrium point $(u^\sharp, \lambda^\sharp)$ of this algorithm solves the system of Kuhn-Tucker optimality conditions of problem (1): for all positive ε and ρ,

$$u^\sharp = \Pi_{U^{\mathrm{ad}}}\Big(u^\sharp - \varepsilon\big(\nabla_u J(u^\sharp) + \nabla_u \Theta(u^\sharp)\,\lambda^\sharp\big)\Big)\,, \tag{31a}$$

$$\lambda^\sharp = \Pi_+\Big(\lambda^\sharp + \rho\big(\Theta(u^\sharp) - \alpha\big)\Big)\,. \tag{31b}$$

We will write algorithm (6) (with ρ^k proportional to ε^k) more compactly: we set $x = (u,\lambda)$ and write

$$x^{k+1} = \Pi(x^k - \varepsilon^k\,\psi^k)\,, \tag{32}$$

where Π stands for the projection operation on $U^{\rm ad} \times \mathbb{R}^d_+$ and ψ^k is driven by an underlying process of i.i.d. drawings ξ^{k+1}, independent of $\{x^i\}_{i\leq k}$. Let $\mathcal{F}^k$ be the filtration generated by $\{x^k, \{\xi^i\}_{i\leq k}\}$ so that ψ^k and x^{k+1} are $\mathcal{F}^{k+1}$ measurable.

With the stochastic estimates produced by the AC and FD techniques considered so far in this paper, we obtained *biased* estimates of $\nabla_u \Theta$ (and the bias sometimes also affects the estimate of Θ itself), with a bias going to 0 as $k \to +\infty$. We will denote $\Psi(x^k)$ the correct value of the vector field at x^k, namely

$$\nabla J(u) + \nabla \Theta(u)\lambda\,, \tag{33a}$$
$$\alpha - \Theta(u)\,, \tag{33b}$$

that with which an equilibrium point satisfies (see (31)):

$$x^\sharp = \Pi\big(x^\sharp - \varepsilon \Psi(x^\sharp)\big) \tag{34}$$

for all positive ε.

Define the martingale difference ΔM^k, the bias B^k and the variance V^k of $\{\psi^k\}$ by:

$$\Delta M^k = \psi^k - \mathbb{E}(\psi^k \mid \mathcal{F}^k)\,, \tag{35a}$$
$$B^k = \mathbb{E}(\psi^k \mid \mathcal{F}^k) - \Psi(x^k)\,, \tag{35b}$$
$$V^k = \mathbb{E}\|\psi^k - \mathbb{E}(\psi^k \mid \mathcal{F}^k)\|^2\,. \tag{35c}$$

We will use references [10] and [11] in which convergence results and convergence rates of algorithm (32) are provided. Essentially, if the nonlinear projection operation at the r.h.s. of (32) is missing, under conditions on the quantities (35) in connection with the step size ε^k that we will recall below, the trajectory produced by (32) behaves a.s. as that of the deterministic ODE:

$$\dot{x} = -\Psi(x)\,. \tag{36a}$$

In the presence of the projection onto a closed convex set, the differential equation is more complex to write since it involves another process z in which z takes values in the orthogonal cone $C(x)$ to the convex set at the current point x (hence this process effectively appears only at the border of the convex set). The ODE now reads

$$\dot{x} = -\Psi(x) - z\,, \quad z \in C(x)\,. \tag{36b}$$

The role of z is to maintain x in the convex set, as it is the case for x^k produced by (32). It is defined as the "minimum force" which achieves this goal.

5.2 Convergence

In this subsection, we recall the conditions which ensure that the stochastic Arrow-Hurwicz algorithm will behave as its ODE (36) and we refer to the previous subsection to deduce that primal iterates u^k will converge, at least locally, towards the solution $u^\sharp$ (assumed unique) of the constrained optimization problem. We then apply those results to the case of biased gradient estimates provided by AC and FD methods to derive a policy on how to tune the parameters r (see (19)) and c (see (26)) as functions of the iteration index k in order to satisfy the convergence conditions.

Lemma 8. *Consider the iteration* (32) *and assume that*

$$\sum_k \varepsilon^k = +\infty\,, \tag{37a}$$

$$\sum_k \varepsilon^k \|B^k\| < \infty \quad a.s., \tag{37b}$$

$$\sum_k (\varepsilon^k)^2\, V_k < \infty\,. \tag{37c}$$

Then, a.s., x^k has the same asymptotic behavior as the solution of (36).

This result follows from [10, Chap. 5].

Proposition 9. *Consider the case when the estimate* (19) *(and possibly* (18) *too) is (are) used in the stochastic algorithm* (6) *(with ρ^k proportional to ε^k) with the following choices of the stepsize ε^k and of the "mollifier" parameter r^k:*

$$\varepsilon^k = k^{-\gamma}, \quad r^k = k^{-\beta/2}\,, \tag{38}$$

for β and γ positive. Then the conditions of Lemma 8 are satisfied if

$$\gamma \le 1, \quad \beta + \gamma > 1, \quad 2\gamma - \beta/2 > 1\,. \tag{39}$$

Proof. The first condition (39) is required by (37a). Theorem 2 states that the bias B^k of AC estimates is in $\mathrm{O}\big((r^k)^2\big) = \mathrm{O}\big(k^{-\beta}\big)$, hence $\varepsilon^k\|B^k\| = O\big(k^{-(\beta+\gamma)}\big)$; therefore (37b) is satisfied under the second condition (39). As for the variance V^k, it is in $\mathrm{O}\big((r^k)^{-1}\big) = \mathrm{O}\big(k^{\beta/2}\big)$ which yields $(\varepsilon^k)^2\, V^k = \mathrm{O}\big(k^{\beta/2-2\gamma}\big)$; thus (37c) is satisfied under the third condition (39). □

Proposition 10. *Consider the case when the estimate* (26) *is used in* (6a) *(with ρ^k in* (6b) *proportional to ε^k) with the following choices of the stepsize ε^k and of the FD parameter c^k:*

$$\varepsilon^k = k^{-\gamma}, \quad c^k = k^{-\beta/2}\,, \tag{40}$$

for β and γ positive. Then the conditions of Lemma 8 are satisfied if, in addition of assumptions **(H1)** *and* **(H2)** *of Theorem 6, one has that*

$$\gamma \le 1, \quad \beta + \gamma > 1, \quad \begin{cases} 2\gamma - \beta/2 > 1 & \text{if } \textbf{(H3)} \text{ is satisfied in Theorem 6,} \\ 2\gamma - 3\beta/4 > 1 & \text{if } \textbf{(H4)} \text{ is satisfied in Theorem 6,} \\ 2\gamma - \beta > 1 & \text{otherwise.} \end{cases} \tag{41}$$

The proof follows the same pattern as previously using the evaluations of Theorem 6, the only changes concerning V^k.

5.3 Convergence Rate

Let β and δ be the integers such that:

$$B^k = \mathrm{O}(k^{-\beta}), \quad V^k = \mathrm{O}(k^{-\delta})\,. \tag{42}$$

Reference [11] provides a comprehensive analysis of the convergence rates of algorithms of type (32) under the following assumption: close to its unique equilibrium point $x^\sharp$ (supposed to lie in

the interior of the convex set onto which Π is the projection), function Ψ admits the following representation:

$$\Psi(x) = A(x - x^\sharp) + \mathrm{O}(\|x - x^\sharp\|^2)\,, \tag{43}$$

where A is a matrix with eigenvalues μ satisfying

$$\overline{\mu} = \min\big(\mathrm{Re}(\mu)\big) > \begin{cases} 0 & \text{if } \gamma < 1, \\ \max\big(\beta, (1+\delta)/2\big) & \text{if } \gamma = 1. \end{cases} \tag{44}$$

Then, a direct application of Theorem 3.1 in [11] gives the asymptotic mean square error (MSE) as a function of the algorithm parameters γ, β, δ and it states that:

$$\mathbb{E}(x^k - x^\sharp)^2 = \mathrm{O}(k^{-\kappa}), \quad \kappa = \min(2\beta, \gamma + \delta). \tag{45}$$

Some comments are in order here regarding the application of this result to our situation. First, the authors of [11] state than when $x^\sharp$ lies on the boundary of the feasible convex set, other techniques (e.g. large deviations) are required to establish convergence rates. In our case, we expect that the probability constraint is active at the optimum, hence the optimal dual variable should be strictly positive. In the example of §3.2, we also have positivity constraints on primal variables and that on u (the first primal component) is active at the optimum (see (14)). But it is felt that the projection is rather helpful in accelerating convergence for this component (see numerical results in the next section). We may consider that, asymptotically, u^k is "frozen" at 0 and does not participate to the dynamics of the algorithm ultimately.

Second, condition (44) may not be satisfied. We will come back on this point in the next subsection. Nevertheless, we used the results of [11] as guidelines for the choice of parameters β and γ to drive the primal solution to its equilibrium in the most efficient way.

That said, in order to achieve the fastest convergence rate, one should seek to maximize κ in (45) over the feasible set defined by (39) or (41) and the expression of δ as a function of β. For the case of AC estimates, $\delta = -\beta/2$, the minimum of 2β and $\gamma - \beta/2$ is obtained when those two functions are equal, which yields $\beta = 2\gamma/5$ and a value of $4\gamma/5$; because of the first condition (39), the maximal possible value is obtained with $\gamma = 1$, which yields $\beta = 2/5$ and $\kappa = 4/5$, and we check that this pair (β, γ) satisfies all conditions in (39). Observe that our heuristic reasoning at the end of §4.1.1 and §4.1.2 in order to tune the parameter r when N i.i.d. samples are available (here N is the iteration index k) yields the same results (see (25) in particular).

For the case of FD estimates, under assumption **(H3)** of Theorem 6, the calculations and conclusions are the same. Under assumption **(H4)**, $\delta = -3\beta/4$ and the optimal values are $\beta = 4/11$, $\gamma = 1$, $\kappa = 8/11$ which is of course worse than the previous case. Finally, in the worst case for FD, we get $\beta = 1/3$, $\gamma = 1$, $\kappa = 2/3$.

The following result is a direct application of [11, Th. 4.1 and 4.2]. This CLT gives additional information on the asymptotic behavior of the iterates of (32).

Theorem 11. *Consider algorithm* (32) *with assumptions* (43), (42) *and* $\varepsilon^k = 1/k$ *(that is,* $\gamma = 1$ *in* (38) *or* (40)*). Let*

$$X^k = k^{\kappa/2}(x^k - x^\sharp)\,,$$

with κ *as in* (45)*. If* $2\beta \geq 1 + \delta$*, then as* $k \to \infty$*,* $X^k - k^{\kappa/2-\beta} H_b \bar{B}$ *converges in distribution*

towards a normal distribution of mean 0 and covariance Σ where:

$$\begin{aligned}
&\bar{B} = \lim_{k\to\infty} k^\beta B^k\,,\\
&H_b = A - \beta I\,,\\
&H = A - \big((1+\delta)/2\big) I\,,\\
&R = \lim_{k\to\infty} k^\delta\, \mathbb{E}(\Delta M^k (\Delta M^k)^\top \mid \mathcal{F}^k)\,,\\
&\Sigma H + H^\top \Sigma = R\,,
\end{aligned}$$

where $^\top$ denotes transposition.

Remark 12. From the definition of H_b above and the appearance of $A - \big((1+\delta)/2\big)I$ in the definition of Σ, it is apparent that the strong stability condition (44) (case $\gamma = 1$) ensures that both these matrices are positive definite, so that Σ is well defined. Indeed, with our choices, H_b and H are equal.

5.4 The Case of Arrow-Hurwicz Algorithms

We know discuss the properties of matrix A in the situation of Arrow-Hurwicz algorithms. This matrix has been introduced in (43) in general, and the operator Ψ is defined by (33) in our case. Thus, A is the linearized version of that Ψ at the equilibrium point $x^\sharp$, that is,

$$A = \begin{pmatrix} \frac{\partial^2 L(u^\sharp,\lambda^\sharp)}{\partial u^2} & \frac{\partial^2 L(u^\sharp,\lambda^\sharp)}{\partial u\,\partial \lambda} \\ -\frac{\partial^2 L(u^\sharp,\lambda^\sharp)}{\partial \lambda\,\partial u} & -\frac{\partial^2 L(u^\sharp,\lambda^\sharp)}{\partial \lambda^2} \end{pmatrix} = \begin{pmatrix} J''(u^\sharp) + \big(\lambda^\sharp\big)^\top \Theta''(u^\sharp) & \big(\Theta'(u^\sharp)\big)^\top \\ -\Theta'(u^\sharp) & 0 \end{pmatrix} \tag{46}$$

However, among the constraints Θ, only those saturated (that is, satisfied with equality) at the equilibrium point should be taken into account together with their corresponding multipliers (that is, the non saturated constraints are virtually absent asymptotically).

Under the assumptions that the gradients of saturated constraints are linearly independent (or, otherwise stated, the operator in the upper right-hand corner of the matrix is injective), and that the Hessian of the Lagrangian (that is, the operator in the upper left-hand corner) is positive definite, it can easily be proved that the real part of the eigenvalues of A are positive (see [2, proof of Proposition 4.4.2]). This is condition (44) in the case $\gamma < 1$. When $\gamma = 1$, condition (44) is stronger and will be discussed shortly in the case of our example. Observe that if we assume that the only saturated dualized constraint is the probability constraint (which is the case in our example), then we should assume that the gradient of this probability function at the equilibrium is not zero.

Going back to example of §3.2, matrix A (restricted to the variables (u, v, λ_2)) is equal to

$$\begin{pmatrix} 0.944 & 1.002 & -0.621 \\ 1.002 & 1.211 & -1.181 \\ 0.621 & 1.181 & 0 \end{pmatrix}$$

with eigenvalues $0.974 \pm 0.753\ i$ and 0.207. As predicted, the real parts are positive but the smallest one is equal to 0.207 which is *not* greater than 2/5. Thus, condition (44) (case $\gamma = 1$) is not satisfied (with $\beta = (1+\delta)/2 = 2/5$). However, in the same way as we ignored multipliers corresponding to non saturated constraints because they are stuck to 0 asymptotically, we may consider that the part u of primal variables is "out of the game" ultimately because u is stuck to 0 (the constraint $u \geq 0$ is saturated) at the end of the transient part of the algorithm (remember that the ODE (36a) is to be replaced by the more complex dynamics (36b) when following boundaries of the admissible domain). Therefore, we consider a reduced matrix A by keeping only the 2×2 lower right-hand block (corresponding to the pair (v, λ_2)). The eigenvalues of this reduced matrix are $0.605 \pm 1.014\ i$ and now condition (44) is satisfied even for the case $\gamma = 1$.

6 Numerical Results

Algorithm (6) has been used to solve the example of §3.2 with the AC and FD estimates.

$$u^{k+1} = \Pi_{U^{\rm ad}}\Big(u^k - \varepsilon^k\big(\nabla_u j(u^k,\xi^{k+1}) - \widehat{\nabla_u P}(u^k,\xi^{k+1})\,\lambda^k\big)\Big)\,, \tag{47a}$$

$$\lambda^{k+1} = \Pi_{+}\Big(\lambda^k + \rho^k\big(\pi - \widehat{P}(u^{k+1},\xi^{k+1})\big)\Big)\,. \tag{47b}$$

More precisely, for the AC method, $\widehat{\nabla_u P}(u,\xi)$ should be interpreted as the gradient estimate (19), applied to the example (see (24)); $\widehat{P}(u,\xi)$ is either $p(u,\xi)$ as in (21) or the biased estimate given by (18) (see Remark 1). We tested both versions numerically and there was no significant difference. The estimate (18) was retained for the rest of experiments. Of course, parameter r^k is adjusted according to the rule $r^k = ak^{-1/5}$ where a is a positive constant to be tuned.

For the FD method, $\widehat{\nabla_u P}(u,\xi)$ is given by (26), applied to the example. Again, parameter c^k is adjusted as $c^k = bk^{-1/5}$ where b is a positive constant to be tuned. For $\widehat{P}(u,\xi)$, we used (21).

Numerical experiments are performed according to the following protocol:

- all runs of the algorithms start from the same initial conditions:

  $$u^0 = 0.2, \quad v^0 = 0.8, \quad \lambda_1^0 = 0.5, \quad \lambda_2^0 = 0.3\,.$$

  Recall that the solution is given by (14) and all results will be expressed in terms of differences with those optimal values (hence the equilibrium point for all variables is at 0).

- For AC and FD, 100 runs of the algorithms are performed using the same 100 sequences of pseudo-random numbers to generate Monte Carlo samples of ξ according to the distribution of this variable.

- 5000 thousands iterations are performed for each run.

- For AC and FD, averages of the differences $x^k - x^\sharp$ are computed over the 100 runs together with their standard deviations. What will be shown on the plots are the trajectories of the "average ± standard deviation" of those quantities as functions of the iteration index k.

- The parameters a, b, d, e, f, g appearing in the following rules:

  $$r^k = \frac{a}{k^{1/5}}, \quad c^k = \frac{b}{k^{1/5}}, \quad \varepsilon^k = \frac{d}{e+k}, \quad \rho^k = \frac{f}{g+k},$$

  are tuned by some trials to try to obtain the "best" results for both methods.

Figure 6 shows the plots for the four variables and for the AC (continuous line) and DF (dotted line) methods. Again what is displayed is the "average ± standard deviation" over 100 runs. Results obtained on this example are very close (with maybe a slight advantage to FD in the earliest iterations) with both methods. This confirm the estimation of variance and bias made with *Mathematica* around the optimum for the estimates obtained with the two methods.

7 Conclusions

This paper discussed the problem of stochastic optimization under probability constraints and in particular methods for solving them numerically. Although there exist other ways of taking care of risk considerations in decision problems under uncertainty, we discussed the fact

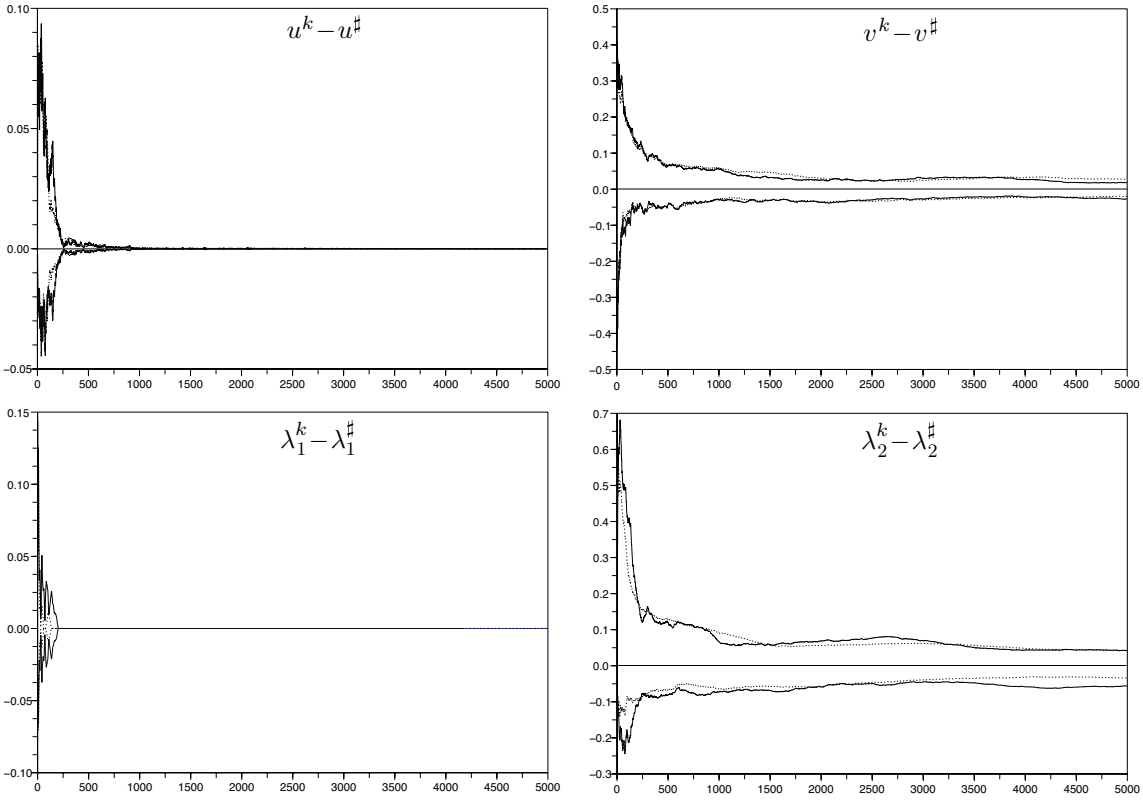

Figure 6: Average ± standard deviation for AC (solid line) and FD (dotted line) algorithms

(§1.2) that probability constraints are sometimes the most straightforward way of expressing and quantifying risk in some circumstances.

Unfortunately, as shown by the discussion and examples in §3, probability constraints may be the source of several pathologies, and the loss of convexity is the most frequent one. Nevertheless, one must address the problem of numerical resolution with approaches which may fail in the worst cases but which may also succeed to solve nontrivial problems. Our strategy is based on duality and stochastic gradient algorithms. Duality, and the use of stochastic Arrow-Hurwicz algorithms, require the existence of a saddle point of the Lagrangian, which is not granted for the reasons advocated above. The use of augmented Lagrangians would certainly increase the chance of existence of saddle points but, in combination with stochastic algorithms, it raises new difficulties (namely, the operator of mathematical expectation would appear inside a nonlinear function). This new topic will be addressed in a forthcoming paper.

Apart from this problem of saddle point existence, the search of this saddle point by stochastic gradient algorithms is made possible by expressing the probability constraint as an expectation involving a discontinuous function. In this paper, we proposed two ways to overcome this difficulty, and we studied the convergence and convergence rate of the resulting algorithms. The two methods provide *biased* stochastic estimates of the constraint gradient. Although their implementation on a simple example showed a similar behavior, the theoretical results reveal that in more general situations, the "mollifier" (or "Approximation by Convolution" — AC) method should be of more general use and robustness then the "Finite Difference" (FD) method. We defer to a forthcoming paper to propose other estimation techniques providing *unbiased* estimates and based on techniques of integration by parts.

Still, the surface of this difficult field of numerical resolution of probability constrained

stochastic optimization problems has been just scratched here, and several directions remain open for future investigations. For example, we have considered here only events (whose probability is constrained) which are described only by a scalar constraint and the case of events described by multidimensional constraints may raise new questions (although the techniques discussed in the present paper seem ready for an extension to this case).

References

[1] Arrow K, Hurwicz L, Uzawa H (1958)*Studies in nonlinear programming*. Stanford University Press, Stanford, CA

[2] Bertsekas DP (1999) *Nonlinear Programming*. 2nd Edition, Athena Scientific, Belmont, Mass.

[3] Cohen G (1980) Auxiliary problem principle and decomposition of optimization problems, *J. Optim. Th. Appl.* 32 (3)

[4] Cohen G, Zhu DL (1984) Decomposition coordination methods in large scale optimization problems. The nondifferentiable case and the use of augmented Lagrangians. In: Cruz JB (ed.) *Advances in Large Scale Systems, Vol. I.* JAI Press, Greenwich, Conn., pp. 203–266

[5] Culioli JC, Cohen G (1990) Decomposition-coordination algorithms in stochastic optimization. *SIAM Journal of Control and Optimization* 28:1372–1403

[6] Culioli JC, Cohen G (1995) Optimisation stochastique sous contraintes en espérance. *C.R. Acad. Sci. Paris*, t. 320, Série I, pp. 753–758

[7] Ermoliev YM, Norkin VI, Wets RJB (1995) The Minimization of Semicontinuous Functions, Mollifier Subgradients. *SIAM J. Contr. Optim.*, 33:149–167

[8] Henrion R (2002) On the Connectedness of Probabilistic Constraint Sets. *J. Optim. Th. Appl.* 112:657–663

[9] Kall P, Wallace SW (1994) *Stochastic Programming*, John Wiley and Sons, Chichester, UK

[10] Kushner H, Yin G (2003) *Stochastic approximation and recursive algorithms*, Stochastic modelling and applied probability, vol. 35, 2nd Ed., Springer Verlag, NY

[11] L'Ecuyer P, Yin G (1997) Budget dependent convergence rate of stochastic approximation. *SIAM Journal on Optimization*, 8:217–247

[12] Prekopa A (1995) *Stochastic Programming*, Math. and Appl. 324, Kluwer

[13] Rockafellar RT, Uryasev S (2001) Conditional Value-at-Risk for General Loss Distributions. *EFA 2001 Barcelona Meetings, EFMA 2001 Lugano Meetings; Univ. of Florida, ISE Dept. Working Paper Nr. 2001-5*

[14] Uryasev S (2000) Introduction to the Theory of Probabilistic Functions and Percentiles (Value-at-Risk). In: Uryasev S (ed.) *Probabilistic Constrained Optimization: Methodology and Applications*, Kluwer, pp. 1–25